Modelling for Natural Language Understanding

Robert BAUD[1,] Christian LOVIS[1], Laurence ALPAY[1], Anne-Marie RASSINOUX[1],
Jean-Raoul SCHERRER[1], Anthony NOWLAN[2], Alan RECTOR[2]
[1] University State Hospital of Geneva, Switzerland
[2] University of Manchester, United Kingdom

Natural Language Understanding (NLU) is a rapidly growing field in medical informatics. Its potential for tomorrow's applications is important. However, it is limited by its ability to ground its components on a solid model of the domain. This opens the way for the emergence of the discipline of medical domain modelling, as part of the vast field of Knowledge Base (KB) engineering.

This article aims at describing the current development of a mutilingual natural language system, strongly oriented towards the semantics of the domain. Special emphasis is presently given to the task of building a domain model, and to establish direct links with the language platform. The result is a model-driven NLU system. Numerous benefits are expected in the long term.

THE NEED FOR MODELLING

Any intelligent program in medical informatics relies essentially on a model of the medical subdomain of concern. However, such a model is rarely explicitly defined, it lies hidden somewhere between the brain of the author and the written programmes themselves. Since the seventies, the new trend has been to make the constraints of a domain explicit. This has led to the so-called rule-based systems, following the epic period of Mycin. Numerous projects have been developed in this line. Promises made in the past have not been kept, and it is agreed today that a surface approach to knowledge representation is precarious. Over simplistic rule-based systems have only the power of assembly language: they are not really a knowledge representation of a problem, but a pseudo procedural approach, which is a dead end. On the contrary, a deep representation is often necessary, especially for the more trivial tasks, as demonstrated by the CyC approach [1], on common sense reasoning. Moreover, an engineering approach is not enough, as emphasized by Doug Lenat saying *"Representation of knowledge is an art, not a science"*.

At first sight, NLU and KB modelling would seem to be independent disciplines. This situation partly results from the numerous developments of parsing techniques, based on formal grammars [2], in the line of Chomsky. Linguists are interested at first in the texts and the related syntax. They will consider the domain and the underlying semantic only if unsolved ambiguities remain. They aim at covering the whole range of human language expressions, and are reluctant to incorporate domain dependent constraints, because they lose generality. On the other hand, KB engineers are dealing with conceptual entities, abstracted from any language. They do not take into consideration morphology, sentence construction, syntax, usage and local jargon; in other words, they miss some essential features which cannot be avoided with medical documents. The gap is there, and the bridge builders are awaited by those who have recognized the potential of a convergent approach.

THE MODELLING PROCESS

Ontology is the backbone of any domain model. First, we need a typology of conceptual entities, linked in the form of a lattice, from the general concepts to the specialized ones, preserving a multiple inheritance scheme. Considerable effort is necessary in the initial design phase of a domain typology, because any misconception at this stage would generate problems in the future.

A complementary facet of this ontology is the tree of relationships, sometimes named slots in frame based systems, or attribute - value pairs. The designer defines the necessary relationships in a domain model. Concepts and relationships are basically the same building objects: they both stand as conceptual entities, because any relationship can be reduced to a concept and the usage of a small set of elementary relationships. Figure 3 gives an example of this situation.

Conceptual entities and relationships are the basic building blocks of semantic networks [3, 4], which are the grounds of our modelling approach. Our current experiment is performed using the semantic network of the GALEN consortium of the AIM

0195-4210/92/$5.00 © 1994 AMIA, Inc.

The physician has shown by laryngo-fibroscopy a paralysis of the left vocal cord.

```
[HUMAN_PROCESS:statement]
    (AGENT) -> [ACTOR:physician]
    (ORIGN) -> [TEST_PROC:fibroscopy]
        (THEME) -> [BODY_PART:larynx]
    (THEME) -> [DISEASE:paralysis]
        (LOC) -> [BODY_PART:vocal_cord]
            (PREC) -> [REGION:left]
```

Figure 1: A typical analysis of a medical sentence, as issued from a discharge letter. The conceptual graphs are based on the NLU typology of concepts and relationships. We can see in this CG that the verb "to show" is nominalized to the noun "statement".

project (Advanced Informatics in Medicine) of the CEC (Commission of the European Communities). A set of tools have been developed by A. Rector [5, 6] and his group, under the name of GRAIL (Galen Representation And Integration Language). On the natural language side, French and English dictionaries and a corresponding language analyser have been produced by our group [7]. Figure 1 shows the level of complexity which is presently handled. The knowledge representation of texts is in the form of conceptual graphs (CG), as presented by J. Sowa [8]. Examples in this paper have been taken from this framework.

In the Galen context, the modelling phase consists of the elaboration of a semantic network, the goal being to accommodate all sensible medical expressions. Two processes are available for this task. First, a sanctioning mechanism allows the specification of what is sensible, i.e. what is not sanctioned is not accepted. Second, an indefeasible definitional schema is ready to grasp the reality in the model. These two mechanisms are shown in figure 2. Any concept in the network is later made available to the outside world as a canonical form. It is a knowledge representation of a sensible medical expression, which may be found in numerous medical documents.

The sanction of a sensible expression:

```
Hepatitis triple hasCause Virus
        possible.
Fracture triple hasLocation Bone
        necessary.
```

The definitional assertion of a conceptual entity:

```
Cancer which <hasLocation Lung has-
    CellType EpithelialCell>.
(Doctor which hasRole PrivateActor)
    name GeneralPractitioner.
```

Figure 2: The two processes used when building a model: the "triple" defines the constraints (what is allowed) and the "which" introduces the definition of a new concept, as a specialisation of existing concepts.

CONCEPTUAL GRAPHS AND GRAIL

When comparing a canonical form from the Galen language GRAIL with a conceptual graph as used in the NLU analyser (see figure 5), it appears immediately that they are both vehicles for knowledge, based on first order logic (as shown in [9]), with converging expressiveness. This is also true, when comparing GRAIL with frame based systems, which contain slot - value pairs, despite the fact that such systems may not have the flexibilty of semantic networks. This paves the way for a semi-automatic translation from one representation to another. The benefit is a potentialisation of the model, having access to a natural language input (analyser) and output (generator), as well as potentialisation of the NLU system, being grounded on a solid source of knowledge, for the task of generating its dictionaries and concept definitions.

FROM SEMANTIC NETWORK TO NATURAL LANGUAGE

The subject of this section is to consider six possible bridges, linking a domain model to a NLU system.

1. Ontological definitions

Ontological definitions can be realized in numerous ways, but a number of them lead to narrow issues, and have sooner or later to be discarded. Only the modelling approach garantees well grounded solutions. Strictly speaking, on the NLU side, less constraints are present when defining the hierarchy of concepts and relationships, because the text itself is considered to be self consistent.

The problem of specificity of relationships, which is unlimited on the GRAIL side, as mentioned above, is solved with the adjunction of a "reldef" feature on the NLU side. Any new relationship in the model must now be defined explicitly, starting from the set of basic relationships (actually no more than 60 such relationships have been defined). Figure 3 illustrates the handling of relationships. Following this scheme, any programme handling CGs, is able to perform expansions and contractions of relationships when needed, or upon the user's request.

In the GRAIL language:

```
AnyConcept which hasColor Red
```

On the NLU side we have:

reldef `hasColor(x,y)` **is**

```
[x]->(CHRC)->[COLOR]->(VALUE)->[y]
```

giving the final CG:

```
[ANYCONCEPT]
   -> (CHRC) -> [COLOR]
      -> VALUE) -> [COLOR_VALUE:red]
```

Figure 3: Demonstration of the handling of a newly defined relationship like "hasColor", and its reduction to two relationships from the basic set like "CHRC" and "VALUE".

2. Populating the lexicons

Populating the dictionaries is of great importance. A given medical specialty would require up to 20'000 words. The entire domain of medicine is estimated at more than 200'000 words. Such an estimation is based on the recent publication of SNOMED International [9], from which we have extracted the following values: 12'385 topography terms, 4'991 morphology terms, 16'352 function terms, 28'622 disease and diagnostic terms and 27'033 procedure terms. Other authors are speaking of 300'000 medical words [10]. This means that a huge task has to be considered, with adequate manpower resources. Even under the best conditions, computer-aided incorporation of the basic vocabulary is mandatory, as well as strong validation processes. It has long been recognized that the task of generating large dictionaries is extremely time consuming, and could be a reason for project failure, without adequate tools and manpower resources. This fact is even more important in a multilingual environment.

English version

```
colotomy,
noun(no,sing),
[SURGICAL_TREATMENT:incision]
   (THEME) -> [ORGAN:colon]
```

French version

```
colotomie,
nom(fem,sing),
[TRAITEMENT_CHIR:incision]
   (THEME) -> [ORGANE:colon]
```

Figure 4: representation of a concept in two natural languages, English and French, with a common semantic representation.

The following GRAIL canonical form:

```
(RespFinding which <
hasState Accelerated
isTypeOf MeasureThing
hasFunction Ventilation>)
name Tachypnea
```

is translated to the conceptual graph:

```
[FINDING: tachypnea]
   (STATE) -> [Accelerated]
   (TYPE) -> [Measurement]
   (FUNCT) -> [Ventilation]
```

Figure 5: This example shows the correspondance between a GRAIL canonical form and its CG counterpart. Automatic translation has been effectuated.

It is well-known that the general relationship between a concept and the different terms pertaining to different languages, is one to many. This means that a number of different words from different syntactical categories, are candidates for the expression of a single concept. Different languages will lead to different words, and different syntaxes, but all represent the same concept. All the words originating from the same concept would have the same common semantical representation. This is the potential benefit of the domain model: it would enable a dictionary of concepts to be populated, from which the language lexicons are compiled.

To illustrate this process, figure 4 shows the dictionary entries in English and in French for the concept "colotomy". Such a concept, for non-medical experts in digestive surgery, can be better understood when reading its CG representation.

The model builder is not the lexicon builder. Two different tasks must coexist in order to be suitably developed. However, it is obvious that they are directly linked to each other, and this must be enforced by adequate tools. Any newly introduced concept in the domain model should automatically generate a corresponding entry in the lexicon, to be completed manually later, for the collection of expressive words of this concept, and for the syntactical part.

3. Conceptual definitions and properties

The model of a medical domain contains much more knowledge information than what is strictly necessary for the task of analysing free texts. The model essentially consists of three parts: first, the ontology; second, a sanctioning mechanism; and third, an assertional mechanism. This last aspect is used for two different purposes: 1) to enter explicit definition in addition to the definitional role of the ontology; 2) to grasp properties and contextual information for each concept. This represents the largest source of knowledge for which the modelling process is extremely promising.

When analysing sentences, the necessary information to understand the meaning is not always present in the text. In fact, a lot of information is implicitly defined through the context or by default. This knowledge gap can be filled by the user quite easily. However, a computer program only knows what it has been fed with, and nothing more. The modelling of the properties belonging to the conceptual entities, is a first answer to this non trivial task. A model has basically two roles: the definition of entities and the collection of their characteristics as a set of properties. This set is the recipient where a number of probable, possible, default, feasible values are assembled, being the representation of the context of this entity.

This aspect is even stronger when querying a data base of text representation. The query process is certainly the major goal of a NLU system. Between a query and a set of texts, inference capabilities are essential for a successful match.

4. Semantical compatibility rules

The method selected for natural language analysis is known as "Proximity processing" and has been described in other papers [7, 11, 12]. It is mainly based on the semantical information from the underlying domain. This method takes advantage of the properties of immediate constituents in sentences, in order to build meaningful terms. For this, a limited set of formal mapping rules, and mixed syntactical and semantical information, are required.

The rules are more or less equivalent to the sanctions as expressed in the GRAIL model. However, the sanctioning system has certainly to be stronger at the level of the model than at the level of the text analyser, because the model has to be concise, and able to precisely recognize sensible expressions. As far as NLU is concerned, it seems reasonable to consider, a priori, that a text is sensible per se, and the need for sanctioning arises only in the presence of ambiguities. This means that compatibility rules should be sanctioned by the model, and not vice versa. An adequate browsing tool-will be provided to allow an easy consistency check. The example in figure 6 illustrates this kind of validation.

The following NLU rule:

```
rule(57,Syntax,cl_pain,
cl_body-part,'LOC').
```

is deducible of the GRAIL sanction:

```
Pain triple hasLocation BodyPart
possible.
```

Figure 6: Example of an NLU rule (#57), for which arguments 3, 4 and 5 are automatically derived from the "triple" belonging to the model. The second argument "Syntax" is instantiated to a value dependent of the language itself.

FROM NATURAL LANGUAGE TO SEMANTIC NETWORK

5. Validation of expressions

The main advantage of a domain model for NLU is the possibility to check any natural language sentence against the model, and to decide whether or not it is sensible, in relation to the model. This is a way to validate any assertion from the outside world. Such a facility is important in the presence of ambiguous sentences.

This statement is correct only if the model and the analyser are perfect: this is an ideal situation, still far from the current state-of-the-art. When a sentence is found not to be sensible, this may be also interpreted as a failure of the analyser to solve some ambiguities, or any other misunderstanding of the content of the sentence. Alternatively, the sentence may bring new knowledge, not yet incorporated into the model. In the latter situation, the question arises: how to accept natural language sentences as modelling information?

This is a true challenge for NLU: to reverse the process where the modelling phase precedes the natural language processing phase, and to grasp additional knowledge from the unlimited corpus of written documents. This is presently a vision, far from the present capabilities. In order for natural language to take in additional knowledge, it is necessary to build a deep and robust model of the domain, with good coherence and stability, and which is able to resist inconsistencies and ambiguities.

6. Validation of the domain model

Points 1 to 4 are the exchange of knowledge from the model to the NLU system, and point 5 is the validation of NLU sentences by the model. The last, but not the least, important bridge is the feedback loop from the NLU system to the model: the quality of free text handling is strongly dependent on the quality of the model.

Therefore, the NLU system acts as a first class quality assurance process.

STATUS OF WORK

The NLU system discribed here is functioning today in French and in English on the basis of a limited dictionnary of 2000 words. An implementation in German is underway. This multilingal analyser is able to handle satisfactorily noun and verb phrases, but not relative phrases; coordinations and references have not yet been completly solved, and should be improved. Other members of our development team are working on a language generator and a natural language query processor. The GRAIL modelling system is available today, and a number of specific models have been developed.

The bridge between the two systems is presently in a design phase, from which the ideas in this paper are extracted. A pilot program for the alignment of NLU and GRAIL typologies has been achieved, and has shown that such a task is feasible. However, this should be very carefully designed.

Short term goals are: alignment of typologies and implementation of model-based dictionary building tools. Long term goals are: transfer of definitional assertions from the model to the NLU system, automatic validation of NLU constraint rules by the model sanctioning mechanism, and ability to enter new modelling information as free text.

CONCLUSIONS

The potentialisation of two medical informatics fields, one by the other, is certainly a way to new progress. Modelling a medical domain, and natural language understanding of free texts in the same domain, are complementary disciplines.

Current developments in these areas have already been fruitful. NLU is built for robust solutions, including analysis, generation, translation and query processing of free texts. In the short term, the modelling process is legitimated by its use for enforcing NLU; in the long term, reversing the process of the feeding in of knowledge - from free text to the model - is a major step in the general process of knowledge based engineering.

Acknowledgments

The current developments, carried out in the University Hospital of Geneva, are part of the GALEN project (A2012) and the HELIOS project (A2015), and have been made possible thanks to the funding of the AIM programme of the CEC, for which the Swiss government (CERC) has given full support.

The GRAIL language and the modelling methodology have been developed within the framework of the GALEN project, at the University of Manchester.

References

[1] D. B. Lenat, R. V. Guha, Building Large Knowledge-Based Systems: Representation and Inference in the Cyc Project, Addison-Wesley Publishing Company, 1989.

[2] M. King, Parsing Natural Language, London, Academic Press, 1983.

[3] M Minsky, A Framework for representing Knowledge. In: The Psychology of Computer Vision, Mc Graw, Hill, 1975.

[4] J F Sowa, Principle of Semantic Networks: Explanations in the Representation of Knowledge, Morgan Kaufmann Publishers, 1988.

[5] A L Rector, W A Nowlan, S Kay, Conceptual Knowledge: The Core of Medical Information Systems, MEDINFO 92, K C Lun et al. editors, North Holland, p1420-1426.

[6] A L Rector, the GALEN team, The Master Notation, Deliverable no 6 to the CEC of GALEN, internal document, March 1993.

[7] A-M Rassinoux, R H Baud, J-R Scherrer, Conceptual Graphs Model Extension for Knowledge Representation of Medical Texts, MEDINFO 92, K C Lun et al. editors, North Holland, p1368-1374.

[8] J F Sowa, Conceptual Structures: Information Processing in Mind and Machine, Addison-Wesley Publishing Company, 1984.

[9] SNOMED International, Introduction, College of American Pathologists, April 1993.

[10] E.R Gabrieli, D.J. Speth, Basic Paradigms for Electronic Patient Records, Journal of Health Information Management Research, Vol. 1, No 1, 1992.

[11] R H Baud, A-M Rassinoux, J-R Scherrer, Natural Language Processing and Semantical Representation of Medical Texts, Meth Inform Med, 2, 1992.

[12] R H Baud, A-M Rassinoux, J-R Scherrer, Natural Language Processing and Medical Records, MEDINFO 92, K C Lun et al. editors, North Holland, p1362-1367.